3. rakenduslingvistika konverents

Tallinn, 25.–26. märts 2004

Teesid

Eesti Keele Instituut
Tallinna Pedagoogikaülikool

Konverentsi pearahastajad: Eesti Keelenõukogu, Haridus- ja Teadusministeerium
Toetajad: Kirjastus TEA, Eesti Keele Sihtasutus
Logo autor ja kaanekujundaja: Martin Metslang
Kodulehekülg: http://www.eki.ee/rakenduslingvistika.html
Toimetajad: Jane Lepasaar, Maria-Maren Sepper

Ühest objektsüsteemide struktureerimise efektiivsest meetodist

Rein Kuusik, Leo Võhandu
TTÜ Informaatikainstituut
kuusik@cc.ttu.ee, leov@staff.ttu.ee

Klassikaline statistika on üha kiiremini kuhjuva infotulva ees vägisi jänni jäämas. Suhteliselt siledalt kulgenud või silutud nähtuste modelleerimise asemele on aiva ägedamalt tunginud NP-keerulised kombinatoorsed meetodid (nn avastusteadus (*discovery science*)). Klassikalised optimeerimismeetodid on asendunud bioloogiliste ja kultuurialgoritmidega, mille võimsus on sadu kordi suurem. Peamine on seejuures asjaolu, et keeruliste nähtuste süvastruktuuri korrapärasid on nende meetodite abil hoopis kergem avastada.

Muude rahvusvahelises kirjanduses esinevate uute meetodite hulgas on kindla koha võitnud ka meie instituudis loodud nn monotoonsete süsteemide meetod. (Mullat, Võhandu, Kuusik jpt).

Selle meetodi põhiidee on ülimalt lihtne:

1. Objektsüsteemi iga objekti jaoks leidub mõõt, mis on nõrgalt monotoonne, st kui muuta süsteemi ühe objekti mõõtu pluss- või miinussuunas, siis teiste objektide mõõdud muutuvad samas suunas või jäävad konstantseks.
2. Süsteemist kõrvaldatakse selle mõõdu järgi nõrgim objekt ja süsteemi mõõdud arvutatakse uuesti. Seda tegevust korratakse seni, kuni kõik objektid on elimineeritud.

Sama protseduuri võib korrata ka objekte kirjeldavate tunnuste puhul. Lõpuks korrastatakse nii objektid kui tunnused elimineerimisjärjekorda ümber. Süsteemi struktuur avaneb lokaalsete maksimumide kohal tükeldusi tehes. Meetodi olemust kirjeldatakse lihtsa näitega. Demonstreeritakse korrastuse seost Hammingi meetrikaga ja saavutatavat ajavõitu.

Meetodi tööaeg on alati polünomiaalne. Saadud struktuuri-tükeldus on kasutatud mõõdu suhtes alati optimaalne.

Esitatud lihtsast põhiideest on arendatud terve meetodite pere, mille abil saab edukalt uurida mitmesuguseid objekt-tunnus- andme-tabeleid, genereerida automaatselt hüpoteese, leida graafide klikke, üldistada tesauruse päringuid katvuse parandamiseks, uurida tekste mitmel eri moel jne. (Meenutame, et graaf on igasuguste suhete kirjeldamiseks universaalselt sobiv mudel)

Ettekandes tuuakse ka tuutornäide hüpoteeside genereerimise kohta.

Lõpuks visandatakse üldistatud töökeskkond ja näidatakse paari kena demopilti keerulise struktuuriga objektsüsteemide olemuse avamise kohta.

Monotoonsed süsteemid ja tekst

Rein Kuusik, Leo Võhandu
TTÜ Informaaatikainstituut
kuusik@cc.ttu.ee, leov@staff.ttu.ee

Tekst on üpris raske objekt analüüsimiseks. Võimalike vaadete hulk tekstile on vist küll piiramatu. Käesoleva konverentsi korraldajad oma kenas käivituskutses viitasid hämmastavalt laiale probleemiringile, mis ootab toetust ja loodab edeneda mitmesuguste klassifitseerimismeetodite kasutamise abil.

Me ei ole teksti uurimisele spetsiaalselt pühendunud, kuid paljude konsultatsioonide käigus ja juhendamisel on siiski mõningaid kontaktkohti tekkinud. Üldised meetodid on spetsiaaluurimustes korraga nii head kui halvad. Head selles mõttes, et metoodika ei sõltu konkreetsest uurimuseesmärgist ning saadud struktuuripilt annab autorite mitmekümneaastase kogemuse põhjal uurijale alati midagi uut. Häda on aga selles, et see uus tuleb alles lahti mõtestada, sest hoolimata tuttavate kirjelduselementide kasutamisest peitub saadud struktuuripiltides alati midagi hämmastavat, ebatraditsioonilist. Iga uus nähtus aga tekitab alati probleeme.

Kõige esimeseks monotoonsete süsteemide (MS) rakenduseks teksti analüüsimisel võib pidada K. Ääremaa tööd juriidilise infosüsteemi JURIOS loomisel. Teise autori aspirandina kasutas ta juriidilise tesaurusega seotud päringusüsteemis stardipäringu automaatseks laiendamiseks süsteemi terminite temaatilise seostatuse mõõdu infot. Näiteks „lahutamine" laieneb loomulikul viisil „mees, naine, laps, abielupool" jne). Sellist laiendamist võiks teha kuitahes kaugele, loomuliku laienduspiiri andis MS-is kasutatav tuuma mõiste. (Mõiste piir).

Huvitav oli ka V. Sarve doktoritöö rahvalaulust, kus MS abil saadud mustrite skeemid tõid selgelt välja laulustiili ajalised muutused.

Teksti uurimise juures on raskendavaks veel asjaolu, et me ei saa nõuda sõnade ega lausete kirjelduse homogeensust (ühetähendus-

likkust ega sama pikkust). See tähendab, et teksti ei saa esitada andmetabelina. Meie õnneks selgub, et seda polegi tarvis. Peaasi, et on täidetud monotoonsuse nõue ja on määratletud täheühendi kui mustri olemus. Kasutades MS-del baseeruvat mustrite avastamise meetodit, saame eraldada kõikvõimalikud tekstis esinevad täheühendid, nn N-grammid nii sõna alguse suhtes kui asukohast sõltumatult. Algoritmid töötavad üpris kiiresti, keskmiselt „musterdatakse" 1 MB teksti (*ca* 400 lk) 3–4 sekundiga.

Ettekandes esitatakse näiteid MS kasutamise kohta teksti analüüsimisel. Huvitav on näiteks uurida ka autori kirjutamisstiili, millele otsa vaatamiseks kasutame ainult kirjavahemärkide ja muude sümbolite rivi. Visandatakse võimalus eesti keele sõnaperede süsteemi loomise toetamiseks jne.

Sõnede jada asemel võime teksti käsitleda lauseehituslikult, sõnatüüpide jadana. Sellele jadale saab otse rakendada monotoonseid mustriuurimise meetodeid. Tuleb luua mustrite skeem, uurida seda lähemalt ja teha vajalikud ja võimalikud üldistused.

Loodud metoodika on rakendatav igasuguste tekstiteisenduste järel.